Estimation of continuous-time interest rate models: a nonparametric approach

Orazio Di Miscia*

1 December 2004

*This research was carried out when the author was visiting the School of Economics and Management at the University of Aarhus. This research has been supported by a Marie Curie Fellowship of the European Community Programme Improving the Human Research Potential and the Socio-Economic Knowledge Base under contract number HPMT-CT-2000-00139. I would like to thank Bent Jesper Christensen and participants at the workshop "Topics in Applied Economics and Finance", University of Aarhus. Only the author is responsible for any omission and mistakes. E-mail *orazio.dimiscia1@bancaintesa.it*

Abstract

This paper presents a general, nonlinear model for term structure interest rate. The approach is the same of Stanton (1997) but it has been extended to a multifactor model. The novel aspect is that rather than choosing the functional specification of the model, the process is generated from the data using approximation methods for multifactor continuous-time Markov processes. In applying this technique to the short and long end of the term structure for a general two-factor diffusion process for interest rates is possible to find some interesting nonlinearity in the interest rate data that are not considered in almost all parametric specifications of term structure interest rate model of the financial literature.

1 Introduction

The asset pricing theory gives us theoretical tools to value a wide array of contingent claims, starting from a continuous-time model for the dynamics of the underlying state variables. In this field one of the most common uses continuous-time models has been in describing the dynamics of the short-term riskless interest rate. Unfortunately, while the theory tells us what to do once we have a model for the underlying variable, it gives us little or no guidance in choosing the right model in the first place. For example, researchers have proposed many different parametric models of short rate dynamics, each attempting to capture particular features of observed interest rate movements. Moreover, empirical tests of these models have yielded mixed results. As a result, several recent researchers have used *nonparametric* techniques to reduce the number of arbitrary parametric restrictions imposed on the underlying process.

This paper apply the fully nonparametric method of Stanton (1997) in order to estimate a multifactor model for term structure interest rates.

Section 2 is a brief overview about continuous-time models and related estimation problems. Section 3 explains some features of existing parametric and nonparametric models. Section 4 derives the same approximations of Stanton (1997) and it is used to understand the philosophy of Stanton's approach. Section 5 extends the Stanton's approach to a multi-factor diffusion process. The last two sections describe data and results with final comments

2 Continuous–time models and some related estimation problems

In asset pricing models, is convenient to represent the underlying state variable(s) (i.e.short-term interest rate) as a continuous-time diffusion process represented by Itô stochastic differential equation:

$$dX_t = \mu(X, t; \theta)dt + \sigma(X, t; \theta)dW_t \tag{1}$$

where $\{W_t, t \geq 0\}$ is a standard Brownian motion. The functions $\mu(\cdot)$ and $\sigma^2(\cdot)$ are respectively the drift (or instantaneous mean) and the diffusion (or instantaneous variance) functions of the process.

In order to estimate (1) when is not possible to solve in an explicitly way the likelihood function, the common approach is to estimate the parameters θ, by applying maximum likelihood to a suitably discretized version of the

continuous-time process. But if the time-step interval between observations does not shrink to zero and simultaneously the sample size doesn't increase, the discretized estimators do not converge to the population parameters even if other consistent estimators exists (so-called "*infill* assumption", see Lo (1988)). The reason is the inappropriateness of applying maximum likelihood estimation to the discretized process when it is the parameters of the continuous-time process that are of interest. This should not be surprising since the likelihood of the discretized process may be viewed as a misspecification of the true likelihood function. All this imply that identification of drift and diffusion functions from discretely sampled data at fixed interval, is impossible in general without impose a restriction on the form of the drift or diffusion function. This is the commonly used strategy to estimate (1): first parameterizing $\mu(\cdot)$ and $\sigma^2(\cdot)$, then discretizing the model in order to estimate the parameters (see Chan, Karolyi, Longstaff, and Sanders (1992) for an empirical example).

But, as pointed above, discretization–based methods implicitly assume that more data means more frequent data on a fixed period of observation (the time interval between the observations must shrink to zero). Even if such data were available, it is likely that market micro-structure problems, such as bid–ask spread, the discreteness of the prices observed, and the irregularity of the intra–day sampling interval, would complicate considerably the analysis of high frequency data compared to daily or weekly data.

Moreover, empirical tests of these parametric version of (1) have yielded mixed results since that financial data exhibit some features that are not taking into account by the different specifications of the drift and diffusion coefficients.

In contrast to this approach, the nonparametric approach[1] solve some drawbacks. It does not require any functional form for the diffusion and drift coefficient: this is particular relevant when the researcher have no theoretical justification for using a particular parametric specification of eq. (1). Moreover some nonparametric methods applied to semiparametric model (e.g. Aït-Sahalia (1996)) do not require that the sampling interval shrink to obtain asymptotic properties for its estimators: this is relevant to avoid micro–structure problems.

The last feature is not valid when nonparametric estimators are applied

[1]There is a common misunderstanding about the use of the word nonparametric. Often the Aït-Sahalia (1996) approach is considered a nonparametric method but it is a nonparametric estimation of the diffusion coefficient in a "semiparametric" model since that the drift coefficient is estimated specifying the form of the function. Instead the Stanton (1997) approach is a "full nonparametric model" since that the nonparametric estimation is used for both drift and diffusion coefficients

within a fully nonparametric models since that they are prone to discretization bias (e.g. Stanton (1997)). But, on the other hand, in those context the researcher is completely free from binding parametric specifications of drift and diffusion coefficients. Moreover, the consistency and asymptotic normality of the estimator can be achieved increasing the sample size n, and for financial time series we have at our disposal very long time series.

All these argumentations constitute a perfect set up to try to estimate a continuous-time stochastic process nonparametrically.

3 Estimating Diffusion Models

3.1 Parametric Models

In estimating the functions μ and σ in eq. (1), the usual approach is first to specify parametric forms for the functions μ and σ, then to estimate the values of the parameters. Given functions μ and σ, the transition density of $X_{t+\Delta}$ at time $t+\Delta$ conditioned to value X_t at time t, $p(\Delta, X_{t+\Delta}|X_t)$ must satisfy the Kolmogorov forward equation (see Oksendal (1995)),

$$\begin{aligned} \frac{\partial p(\Delta, X_{t+\Delta}|X_t)}{\partial \Delta} &= -\frac{\partial}{\partial X_{t+\Delta}}(\mu(X_{t+\Delta;}\theta)p(\Delta, X_{t+\Delta}|X_t)) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial X_{t+\Delta}^2}(\sigma^2(X_{t+\Delta})p(\Delta, X_{t+\Delta}|X_t)) \end{aligned} \tag{2}$$

In principle, for a given parametrization of μ and σ, we can solve equation (2) for the conditional density p as function of the parameters, then use maximum likelihood to estimate the model's parameters (see, for example Lo (1988)). This approach was followed by Pearson and Sun (1994) in estimating the parameters of the CIR interest rate model, using the fact that, under this process, interest rates are conditionally distributed as multiple of a noncentral χ^2 random variable. Unfortunately, in a few cases equation (2) can be solved explicitly; in the other cases the equation can only be solved numerically, making implementation of maximum likelihood extremely inconvenient. To avoid this problem, often the researcher apply the Generalized Method of Moments (GMM) since that it is possible to specify only certain properties of the distribution, rather than the full likelihood function. Other approaches are the simulated method of moments (SMM) of Duffie and Singleton (1993) and the efficient method of moments (EMM) of Gallant and Tauchen (1996). Hansen and Scheinkman (1995) show how to derive analytic moment restrictions from eq. (1) using the infinitesimal generator (see Oksendal (1995)) of X_t, $\mathcal{L}$ defined by

$$\begin{aligned}\mathcal{L}f(x,t) &= \lim_{\tau \downarrow t} \frac{\mathrm{E}(f(X_\tau,\tau)) - f(x,t)}{\tau - t} = \\ &= \frac{\partial f(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x}\mu(x) + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\sigma^2(x)\end{aligned} \tag{3}$$

For example, their first class of moment conditions can be obtained by noting that, if X_t is stationary, $\mathrm{E}[\phi(X-t)]$ must be independent of calendar time for any function ϕ. This implies that its unconditional expected rate of change must be zero:

$$\mathrm{E}[\mathcal{L}\phi(X_t)] = \mathrm{E}[\phi^{'}(X_t)\mu(X_t)] + 1/2\phi^{''}(X_t)\sigma^2(X_t)] = 0 \tag{4}$$

While these moment conditions are less computational intensive than those of Duffie and Singleton (1993) or Gallant and Tauchen (1996), they do not take advantage of all of the information contained in the discretely observed data. An alternative approach is to use GMM with approximate moment conditions. A well–known example is Chan, Karolyi, Longstaff, and Sanders (1992). In estimating their continuous–time interest rate model:

$$dX_t = (\alpha + \beta X_t)dt + \sigma X^{\lambda} dZ_t \tag{5}$$

They use approximate conditional moments of the form

$$\begin{aligned}\mathrm{E}_t(\epsilon_{t+\Delta}) &= 0 &(6)\\ \mathrm{E}_t(\epsilon^2_{t+\Delta}) &= \sigma^2 X_t^{2\lambda}\delta &(7)\\ \text{where } \epsilon_{t+\Delta} &= X_{t+\Delta} - X_t - (\alpha + \beta X_t)\Delta &(8)\end{aligned}$$

While these are only approximately correct, this approach is the simplest of all to implement.

3.2 Nonparametric Methods

One potentially serious problem with any parametric model, particularly when there is no economic reason why we should prefer one functional form over another, is misspecification. The reason is that even if a model fits interest rate movements well in-sample, this does not necessarily imply that it will price securities well. This is because the price today of an interest rate dependent security depends not on the past interest rates, but on the entire distribution of possible future interest rates between today and the maturity

of the security. Fitting historical data well is no guarantee of matching this entire distribution, leading to the possibility of large pricing and hedging errors (see Backus, Foresi, and Zin (1995)). To solve misspecification problems, recent researches have used nonparametric estimation techniques in order to avoid arbitrary functional forms for μ and/or σ. Within the nonparametric framework the approaches are different. For example, Aït-Sahalia (1996) use a nonparametric estimator for the diffusion coefficient after specifying the form of the drift.

3.2.1 The Stanton's approach

The philosophy of the Stanton's approach is in some way similar to that of Aït-Sahalia (1996). Instead of specify the drift and diffusion functions, he finds a relationship between them and some other variable that can be estimated nonparametrically and come back to the determination of the drift and the diffusion. Stanton start with eq. (1) and he rewrite the conditional expectation $\mathrm{E}_t[f(X_{t+\Delta}, t+\Delta)]$ in the form of a Taylor series expansion (see Stanton (1997)):

$$\begin{aligned} \mathrm{E}_t[f(X_{t+\Delta}, t+\Delta)] &= f(X_t,t) + \mathcal{L}f(X_t,t)\Delta + \frac{1}{2}\mathcal{L}^2 f(X_t,t)\Delta^2 + \dots \quad (9) \\ &+ \frac{1}{n!}\mathcal{L}^n f(X_t,t)\Delta^n + O(\Delta^{n+1}) \end{aligned}$$

where $\mathcal{L}$ is the infinitesimal generator of the process $\{X_t\}$ (see eq.(3)). The most common use of eq. (9) is the construction of numerical approximations to the expectation on the left–hand side, given known functions μ and σ. Stanton change the approach since that he doesn't specify the form of the drift and diffusion function. Using a long enough interest rate series he estimates the conditional expectation nonparametrically, given suitable choices of the function f. Eq. (9) is used to construct approximations to μ and σ.

The similarity between this and the Aït–Sahalia's approach is that in either case the authors start with the proposal to avoid the fully parametrizing of μ and/or σ, because in most cases there are no economic motivation for choice a particular form of the drift and the diffusion coefficient. The difference is that Stanton avoid to specify the drift and the diffusion and the researcher is completely free while Aït–Sahalia bind the analysis choosing the form of the drift. This "more freedom" of the Stanton's approach is paid in terms of asymptotic properties of the estimators. This means that the estimators converges to the true μ and σ at a rate Δ^k where Δ is the time

between successive observations, and k is an arbitrary integer. In practice, to achieve the consistency of the estimators, $\Delta \to 0$. As mentioned above this imply some micro–structure problem when the researcher use infra–daily data. But as can be see in Stanton (1997), the approximation is quite good with $\Delta = 0.004$ (daily data).

4 Constructing Approximations to μ and σ

Following the Stanton's approach is possible to rewrite eq.(9) in the following form:

$$\begin{aligned} \mathcal{L}f(X_t,t) &= \frac{1}{\Delta}\mathrm{E}_t[f(X_{t+\Delta},t+\Delta)-f(X_t,t)] \\ &- \frac{1}{2}\mathcal{L}^2 f(X_t,t)\Delta - \frac{1}{3!}\mathcal{L}^3 f(X_t,t)\Delta^2 - \ldots \end{aligned} \tag{10}$$

Ignoring all terms except the first on the right hand side gives us a first order approximation for $\mathcal{L}f$,

$$\mathcal{L}f(X_t,t) = \frac{1}{\Delta}\mathrm{E}_t[f(X_{t+\Delta},t+\Delta)-f(X_t,t)] + O(\Delta) \tag{11}$$

Obviously is possible to construct higher order approximations using simple algebra manipulation to avoid to calculate derivatives of μ and σ (included in $\mathcal{L}^n$), which are themselves unknown. To see this, consider eq. (10) with a time step of 2Δ:

$$\begin{aligned} \mathcal{L}f(X_t,t) &= \frac{1}{2\Delta}\mathrm{E}_t[f(X_{t+2\Delta},t+2\Delta)-f(X_t,t)] \\ &- \frac{1}{2}\mathcal{L}^2 f(X_t,t)2\Delta - \frac{1}{3!}\mathcal{L}^3 f(X_t,t)(2\Delta)^2 - \ldots \end{aligned} \tag{12}$$

Multiplying eq. (10) by 2,and subtracting eq. (12), yields a second order approximation:

$$\begin{aligned} \mathcal{L}f(X_t,t) &= \frac{1}{2\Delta}\{4\mathrm{E}_t[f(X_{t+\Delta},t+\Delta)-f(X_t,t)] - \\ &- \mathrm{E}_t[f(X_{t+2\Delta},t+2\Delta)-f(X_t,t)]\} + O(\Delta^2) \end{aligned} \tag{13}$$

an approximation to $\mathcal{L}f$ in terms of expectations of functions of only observed values of $\{X_t\}$ which converges to the true function at a rate Δ^2 as

$\Delta \to 0$. The process could continue, generating approximations of successively higher order, using a time step of 3Δ, 4Δ etc.

Generalizing eq. (11) and (13) this approach approximate a generic function $g(x,t)$ (the right hand side of each equation) with $\mathcal{L}f(x,t)$

$$\mathcal{L}f(x,t) = g(x,t)$$

Now is necessary to find the right $f(x,t)$

4.1 Approximation of μ

To derive approximations to the drift μ, consider the function

$$f_{(1)}(x,t) \equiv x \tag{14}$$

From the definition of $\mathcal{L}$, we have

$$\mathcal{L}f_{(1)}(x,t) = \mu(x) \tag{15}$$

Substituting successively into eq. (11) and (13) leads to the following approximations for μ:

$$\mu(X_t) = \frac{1}{\Delta}\mathrm{E}_t[X_{t+\Delta} - X_t] + O(\Delta) \tag{16}$$

$$\mu(X_t) = \frac{1}{2\Delta}\{4\mathrm{E}_t[X_{t+\Delta} - X_t] - \mathrm{E}_t[X_{t+2\Delta} - X_t]\} + O(\Delta^2) \tag{17}$$

4.2 Approximation of σ

To construct approximations to the diffusion σ, consider the function

$$f_{(2)}(x,t) \equiv (x - X_t)^2 \tag{18}$$

From the definition of $\mathcal{L}$, we have

$$\mathcal{L}f_{(2)}(x,t) = 2(x - X_t)\mu(x) + \sigma^2(x) \tag{19}$$

and so, letting $x \equiv X_t$

$$\mathcal{L}f_{(2)}(X_t,t) = \sigma^2(x) \tag{20}$$

Substituting in eq. (11) and (13) yields approximations for σ^2

$$\sigma^2(X_t) = \frac{1}{\Delta}\mathrm{E}_t[(X_{t+\Delta} - X_t)^2] + O(\Delta) \quad (21)$$

$$\sigma^2(X_t) = \frac{1}{2\Delta}\{4\mathrm{E}_t[(X_{t+\Delta} - X_t)^2] - \mathrm{E}_t[(X_{t+2\Delta} - X_t)^2]\} + O(\Delta^2) \quad (22)$$

It is possible to replace the terms in $\mathrm{E}_t[(X_{t+j\Delta} - X_t)^2]$ with the conditional variance $\mathrm{Var}_t(X_{t+j\Delta})$ leading to the following set of approximations for $\sigma(X_t)$:

$$\sigma(X_t) = \sqrt{\frac{1}{\Delta}\mathrm{Var}_t(X_{t+\Delta}} + O(\Delta) \quad (23)$$

$$\sigma(X_t) = \sqrt{\frac{1}{2\Delta}[4\mathrm{Var}_t(X_t + \Delta) - \mathrm{Var}_t(X_{t+2\Delta})]} + O(\Delta^2) \quad (24)$$

5 Estimation of a Continuous-Time Multi-Factor Diffusion Process

Most of the literature about term structure interest rate models agree about the inefficiency of single factor model to adequately fit the stylized facts of interest rate. In order to obtain a model that is consistent with the true process underlying the data theoretical studies promote multi-factor bond pricing (see Brennan and Schwartz (1979), Schaefer and Schwartz (1984), Longstaff and Schwartz (1992)). In this section i apply the Stanton's approach to a multivariate setting providing the nonparametric estimation of the drift and volatility functions of multivariate stochastic differential equation.

Under the "usual" assumption of no-arbitrage opportunities and that bond prices are functions of two state variables, R_t and S_t, the stochastic behaviour of these variables are supposed to follow the (jointly) Markov diffusion process:

$$dR_t = \mu_R(R_t, S_t)dt + \sigma_R(R_t, S_t)dZ_t^R \quad (25)$$

$$dS_t = \mu_S(R_t, S_t)dt + \sigma_S(R_t, S_t)dZ_t^S \quad (26)$$

where the drift, volatility and correlation coefficients (i.e. the correlation between Z^R and Z^S) all depend on R_t and S_t. Define the vector $X_t = (R_t, S_t)'$.

Under suitable restriction on μ, σ, and a function f, we can write the conditional expectation $\mathrm{E}_t[f(\mathbf{X}_{t+\Delta}, t + \Delta)]$ in the form of a Taylor series expansion:

$$\begin{aligned} \mathrm{E}_t[f(\mathbf{X}_{t+\Delta}, t+\Delta)] &= f(\mathbf{X}_t, t) + \mathcal{L}f(\mathbf{X}_t, t)\Delta + \frac{1}{2}\mathcal{L}^2 f(\mathbf{X}_t, t)\Delta^2 + \ldots \quad (27) \\ &+ \frac{1}{n!}\mathcal{L}^n f(\mathbf{X}_t, t)\Delta^n + O(\Delta^{n+1}) \end{aligned}$$

which have the same form of eq.(9) but the infinitesimal generator is defined in a multivariate form as[2]:

$$\mathcal{L} = \left(\frac{\partial f(\mathbf{X}_t)}{\partial \mathbf{X}_t}\right)\mu_{\mathbf{X}}(\mathbf{X}_t) + \frac{1}{2}\mathrm{trace}\left[\mathbf{\Sigma}(\mathbf{X}_t)\left(\frac{\partial^2 f(\mathbf{X}_t)}{\partial \mathbf{X}_t \partial \mathbf{X}_t'}\right)\right] \quad (28)$$

where

$$\mathbf{\Sigma}(\mathbf{X}_t) = \begin{pmatrix} \sigma_R^2(R_t, S_t) & \rho(R_t, S_t)\sigma_R(R_t, S_t)\sigma_S(R_t, S_t) \\ \rho(R_t, S_t)\sigma_R(R_t, S_t)\sigma_S(R_t, S_t) & \sigma_S^2(R_t, S_t) \end{pmatrix}$$

Given an appropriately chosen set of functions $f(\cdot)$ and nonparametric estimates of $\mathrm{E}_t[f(\mathbf{X}_{t+\Delta})]$ it is possible to use eq. (27) to construct approximations to the drift, volatility and correlation coefficients (i.e. μ_R, μ_S, ρ, σ_R and σ_S) of the underlying multi-factor, continuous-time diffusion process without impose any functional form. Using the same algebrical manipulation of the univariate case and a time step of length $i\Delta$ (i=1,2,...,N) we obtain:

$$\begin{aligned} \hat{\mathrm{E}}^i(\mathbf{X}_t) &\equiv \frac{1}{i\Delta}\mathrm{E}_t[f(\mathbf{X}_{t+i\Delta}) - f(\mathbf{X}_t)] = \quad (29) \\ &= \mathcal{L}f(\mathbf{X}_t) + \frac{1}{2}\mathcal{L}^2 f(\mathbf{X}_t)(i\Delta) + \ldots + \frac{1}{n!}\mathcal{L}^n f(\mathbf{X}_t)(i\Delta)^{n-1} + O(\Delta^n) \end{aligned}$$

where the index i represent the time step. Eq. (29) show that if we ignore all terms except the first, each of the $\hat{\mathrm{E}}^i$ is a first order approximation to $\mathcal{L}f$:

$$\hat{\mathrm{E}}^i(\mathbf{X}_t) = \mathcal{L}f(\mathbf{X}_t) + O(\Delta) \quad (30)$$

Consider a linear combination of these approximations $\sum_{i=1}^{N}\alpha_i\hat{E}^i(\mathbf{X}_t)$. Eq. (29) becomes[3] :

[2]The dependency of f on t will be suppress for notational convenience, but the reader must keep it in mind

[3]The model consider a number of $\mathcal{L}f$ terms equal to N, the number of the element of the linear combination

$$\sum_{i=1}^{N} \alpha_i \hat{E}^i(\mathbf{X}_t) = \left[\sum_{i=1}^{N} \alpha_i\right] \mathcal{L}f(\mathbf{X}_t) \quad + \quad \frac{1}{2}\left[\sum_{i=1}^{N} \alpha_i i\right] \mathcal{L}^2 f(\mathbf{X}_t)\Delta + \frac{1}{6} \quad (31)$$
$$+ \quad \left[\sum_{i=1}^{N} \alpha_i i^2\right] \mathcal{L}^3 f(\mathbf{X}_t)\Delta^2 + \ldots$$

The idea is to choose the α_i in a way that *the linear combination* is an approximation of $\mathcal{L}f(\cdot)$ of order N.

In order to do this, the weights $\alpha_1, \alpha_2, \ldots, \alpha_N$ must sum to 1. Furthermore, from eq. (31), to eliminate the term of order $O(\Delta)$, the weights must satisfy the equation:

$$\sum_{i=1}^{N} \alpha i = 0$$

More generally, in order to eliminate the term of order $O(\Delta^n)$, the weights must satisfy the equation

$$\sum_{i=1}^{N} \alpha i^n = 0$$

We can write this set of restrictions more compactly in matrix form as

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 3 & \cdots & N \\ 1 & 4 & 9 & \cdots & N^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2^{N-1} & 3^{N-1} & \cdots & N^{N-1} \end{pmatrix} \alpha \equiv V\alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The matrix V is called a Vandermonde matrix, and is invertible for any value of N. We can thus obtain α by calculating

$$\alpha = V^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Substituting α into eq. (31) and using eq. (29), we get the following approximation of the infinitesimal generator of the process $\{\mathbf{X}_t\}$:

$$\mathcal{L}f(\mathbf{X}_t) = \sum_{i=1}^{N} \alpha_i \hat{\mathrm{E}}^i(\mathbf{X}_t) + O(\Delta)$$

To approximate a particular function $g(x)$, we now need merely to find a specific function f satisfying

$$\mathcal{L}f(x) = g(x)$$

For our purposes, consider the functions

$$\begin{aligned} f_{(1)}(R) &\equiv (R - R_t) \\ f_{(2)}(S) &\equiv (S - S_t) \\ f_{(3)}(R) &\equiv (R - R_t)^2 \\ f_{(4)}(S) &\equiv (S - S_t)^2 \\ f_{(5)}(R,S) &\equiv (R - R_t)(S - S_t) \end{aligned} \tag{32}$$

From the definition of $\mathcal{L}$, we have:

$$\begin{aligned} \mathcal{L}f_{(1)}(R) &= \mu_R(R,S) \\ \mathcal{L}f_{(2)}(S) &= \mu_S(R,S) \\ \mathcal{L}f_{(3)}(R) &= 2(R - R_t)\mu_R(R,S) + \sigma_R^2(R,S) \\ \mathcal{L}f_{(4)}(S) &= 2(S - S_t)\mu_S(R,S) + \sigma_S^2(R,S) \\ \mathcal{L}f_{(5)}(R,S) &= (S - S_t)\mu_R(R,S) + (R - R_t)\mu_S(R,S) + \rho(R,S)\sigma_R(R,S)\sigma_S(R,S) \end{aligned}$$

Evaluating these at $R = R_t$, $S = S_t$, we obtain

$$\begin{aligned} \mathcal{L}f_{(1)}(R_t) &= \mu_R(R_t,S_t) \\ \mathcal{L}f_{(2)}(S_t) &= \mu_S(R_t,S_t) \\ \mathcal{L}f_{(3)}(R_t) &= \sigma_R^2(R_t,S_t) \\ \mathcal{L}f_{(4)}(S_t) &= \sigma_S^2(R_t,S_t) \\ \mathcal{L}f_{(5)}(R,S) &= \rho(R_t,S_t)\sigma_R(R_t,S_t)\sigma_S(R_t,S_t) \end{aligned} \tag{33}$$

Using each of these functions in turn as the function f above, we can generate approximations to μ_R, μ_S, σ_R, σ_S and ρ respectively (see appendix C for a formal derivation of approximations of first, second and third order). For example, the third order approximations (taking square roots for σ_R and σ_S) are:

$$\begin{aligned}
\mu_R(R_t,S_t) &= \frac{1}{6\Delta}[18\mathrm{E}_t(R_{t+\Delta}-R_t)-9\mathrm{E}_t(R_{t+2\Delta}-R_t)+2\mathrm{E}_t(R_{t+3\Delta}-R_t)]+O(\Delta^3)\\
\mu_S(R_t,S_t) &= \frac{1}{6\Delta}[18\mathrm{E}_t(S_{t+\Delta}-S_t)-9\mathrm{E}_t(S_{t+2\Delta}-S_t)+2\mathrm{E}_t(S_{t+3\Delta}-S_t)]+O(\Delta^3) \qquad (34)\\
\sigma_R^2(R_t,S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(R_{t+\Delta}-R_t)^2]-9\mathrm{E}_t[(R_{t+2\Delta}-R_t)^2]+2\mathrm{E}_t[(R_{t+3\Delta}-R_t)^2])+O(\Delta^3)\\
\sigma_S(R_t,S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(S_{t+\Delta}-S_t)^2]-9\mathrm{E}_t[(S_{t+2\Delta}-S_t)^2]+2\mathrm{E}_t[(S_{t+3\Delta}-S_t)^2])+O(\Delta^3)\\
\sigma_{RS}(R_t,S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(R_{t+\Delta}-R_t)(S_{t+\Delta}-S_t)]-9\mathrm{E}_t[(R_{t+2\Delta}-R_t)(S_{t+2\Delta}-S_t)])+\\
&+ \quad 2\mathrm{E}_t[(R_{t+3\Delta}-R_t)(S_{t+3\Delta}-S_t)]+O(\Delta^3)
\end{aligned}$$

The approximations of the drift, volatility and correlation coefficients are written in terms of the true first, second and cross moments of multiperiod changes in the two state variables. If the two-factor assumption is appropriate, and a large stationary time-series is available, then these conditional moments can be estimated using appropriate nonparametric methods. In this paper, conditional moments are estimated using a nonparametric regression model(see appendix 7). The conditioning variable are the short and spread rates. All that is required is that these factor span the same space as the true state variables (see Duffie and Kan (1996) for a discussion, in a linear setting, of the conditions under which this is possible).

6 Implementation

6.1 Data Description

The data are the 3 month Treasury Bill and constant maturity treasury yields on the 10 year U.S. Government bond. They are collected from Datastream and cover the period from January 3, 1982 to December 31, 1998, providing roughly 4000 observations.

All the data are quoted as the midpoints between the bid and asked prices at the close of the business day and they are expressed in annualized form.

The 3 month Treasury Bill is the proxy for the short-rate and the spread between the 10 years and the 3 month yield is the proxy for the slope of

term structure. These variables are chosen to coincide with interest rate variables used in other studies (see Litterman and Scheinkman (1991) and Chan, Karolyi, Longstaff, and Sanders (1992) among others).

Figure 1 show the time series of both the short rate and spread. Over the period, the short rate ranges from 2.619% to 10.677%, while the spread varies from -0.266% to 3.788%. This is a signal of distinct periods of low and high interest rates, as well as spread ranges (i.e. different slopes of the term structure interest rates).

Figure 2 presents a scatter plot of the short and spread rate. This graph is particular useful because in this paper, short and spread rate are the conditioning variables. The figure show some holes at the boundary of range, namely at low short rates (i.e. from 2.619-3.5%) and low spread (i.e. from -0.266 and 2%) at high short rate (i.e. from 9.2-10.677%) and low spread (i.e. from -0.266 and 1.5%) and some other holes in regions close to the boundary (ex. short rate in the range 3.5-4.5% and spread between 1 and 2.25%). This means that the researcher should be cautions in interpreting the implied distribution of interest rate conditional on these values for short rate and spread.

Figure 3 show the histograms of the short and spread rate. It's clear that the shapes of empirical distributions is not normal and table 1 and 2 show results that go in the same direction of the visual indication.

Table 5 and 6 show the autocorrelation at different lags of the short and spread daily rates and daily changes. The pattern of the autocorrelation of the daily rates indicate the presence of a unit roots in the time series while the daily changes seem stationary. Tables 7 and 8 confirm the perception showing the augmented Dickey-Fuller nonstationarity test.

The null hypothesis of nonstationarity is rejected at 10% significance level. Because the test is known to have low power, which is the probability of rejecting the null hypothesis when it is not true, a rejection suggests that stationarity of the series is very likely.

6.2 Nonparametric estimation

6.2.1 Marginal and joint nonparametric distribution

Nonparametric methods are particularly useful when the researcher has no previous knowledge or experience about the data generating process of dataset at hand. In this case, before to estimate the model (parametrically or not), a nonparametric estimation of the density can be important in capturing the stylized facts that need explanation and for judging how well a potential model is likely to fit the data. The important elements in nonparametric

estimation are the choice of a smoothing function - the kernel - and a matrix parameter - the window width or bandwidth matrix. Empirical researches found that the choice of the kernel function is not critical and the "optimal" kernel yield only modest improvements in the performance of the density estimator. This is not the case for the bandwidth matrix that play a crucial role for a good "fitting" of the data (see Appendix A).

Figure 4 and 5 show the marginal nonparametric density estimation of the short and spread rate. The bandwidth used is the "Normal reference rule" that is the optimal value when the true density is Gaussian and the kernel function is also Gaussian (see Scott (1992)) .

This 'unrealistic" choice is due to the purely exploratory purpose of density estimation, in order to highlight some stylized fact like the presence of different "humps". The dataset at hand is quite long and it reflects different "regime shift" relative to different economic period and consequently central bank actions. This is one reason for the multi modal empirical distribution. Same deductions from figure 6 and 7 where the bivariate density is estimated.

6.3 Estimation results of the continuous-time multi-factor models

Drift, diffusion and correlation coefficients of eq. (25) and eq. (26) are estimated using the third order approximations (see eq. (34)). The procedure imply the estimation of different nonparametric conditional moments with relative computational problem and computer time spending (see appendix B for details). Estimation results are shown in figure 8-12.

As mentioned above the researcher should treat with caution the results at the boundary of the range, since that there are no available or sufficient data to estimate the conditional moments.

More precisely, it's not possible to obtain reliable estimation in the region relative to low short rate/low spread rate as high short rate/low spread rate. Even in the other two cases (low short rate/high level spread and high short rate/low spread rate) the number of observations are not sufficient for "robust" estimate.

About the diffusion process of the short rate, for fixed values of the short rate between 4% and 9%, figure 8 show a nonlinear behavior along the spread rate.

This feature seems more visible at low level of short rate (between 4% and 5.5%) and at high level of short rate, close to the boundaries of "reliable" surface. Fixing the value of the spread rate between 2% and 3% and moving along the short rate in the range where more observations occur, we find

a substantial linear relation. The surface reveal some slight humps but is not possible to state with certainty the nonlinearity of the drift because may depend on the precision of the optimized bandwidth matrix. In general, the results show that the drift is not perfectly linear, even if it's hard to conclude that these nonlinearities are the proof to reject the estimates of a parametric model with linear drift specification.

An interesting element is that the surface does not exhibit a mean-reverting behaviour and this is compatible with Aït-Sahalia (1996)'s results. On the other hand there are strong theoretical argumentations to believe that interest rate data cannot mimic the path of a random walk.

In figure 9, the heart of the surface show that the volatility is increasing and reach its maximum value when the short and spread rate are high: when the rates are far from the long-run mean, the volatility increase. A possible explanation could be that there is a "real" tendency of the rates, when they are too high or too low, to match the long-run mean and this is in line with the theory of term structure interest rate .

In particular the surface show a deeper inclination when both the short and spread rate are high and the feature is more evident when we move along the spread rate. In some sense, it seems that the volatility increase when the short rate is high.

Figure 10 and 11 show the estimation of the drift and diffusion function for the diffusion process of the spread rate. For the diffusion part the result is similar to that of diffusion coefficient in figure 9, even if the surface seems "oversmoothed". This is due not only to a different time series but also to bandwidth matrix whose entries are larger than those of the bandwidth matrix for the short rate. One reason could be the difficulties to estimate the bandwidth matrix using this kind of data. The drift function show a behavior that seem exactly the contrary (oversmoothing aside) to the drift function of the short rate. One possible interpretation could be the negative correlation between the short and spread rate.

Figure 12 is relative to the correlation coefficient. It is always negative even if close to zero. It seems more negative at high level of the spread rate and high level of the spread rate. In practice when the slope of the term structure increase, in mean, the short rate decrease to come back to the long-run mean.

7 Conclusions

This paper provides a method for estimating in a nonparametric way a multi-factor continuous-time Markov processes using the approach of Stan-

ton (1997). The technique has been applied to the short and long-end of the term structure for a general two-factor, continuous-time diffusion process for interest rates. In estimating this process, the results show some nonlinearities in the drift coefficients.

This seem a surprising result, since that almost all parametric model, in the financial literature about term structure estimation, are specified with a linear drift.

Obviously, this results should be treat with caution since that it could be an artifact of the estimation methods and not a genuine nonlinearity.

Another results is that the volatility of interest rates is increasing in the level of interest rates, only for sharply, upward sloping term structures. Thus, the result of previous studies, suggesting an almost exponential relation between interest rate volatility and levels, is due to the term structure on average being upward sloping, and is not a general result per se.

There are several advantages of the procedure adopted in this paper. First, there is a constant debate between researchers on the relative benefits of using equilibrium versus arbitrage-free models. Here, we circumvent this issue by using actual data to give us the process. Since the real world coincides with the intersection of equilibrium and arbitrage-free models, the model is automatically consistent.

Second, the approach of this paper may be useful in providing forecasts of the conditional distribution of changes in the term structure of interest rates.

Moreover it could be used to evaluate the performance of a parametric model since that within this nonparametric framework no previous specification are needed: the data speak for themselves.

References

Aït-Sahalia, Y. (1996): "Nonparametric Pricing of Interest Rate Derivatives Securities," *Econometrica*, 64, 527–560.

Backus, D. K., S. Foresi, and S. E. Zin (1995): "Arbitrage opportunities in arbitrage free models of bond pricing," Working paper, New York University.

Bowman, A. W. (1984): "An Alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 71, 353–360.

Brennan, M., and E. Schwartz (1979): "A Continuous Time Approach to the Pricing od Bonds," *Journal of Banking and Finance*, 2, 133–155.

Chan, K., F. Karolyi, F. Longstaff, and A. Sanders (1992): "An Empirical Comparison of Alternative Models of the Short–Term Interest Rate," *Journal of Finance*, 47, 1209–1227.

Chu, C., and J. Marron (1991): "Comparison of two bandwidth selectors with dependent errors," *Annals of Statistics*, 19, 1906–1918.

Duffie, D., and R. Kan (1996): "A Yield Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Duffie, D., and K. Singleton (1993): "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929–952.

Fan, J. (1993): "Local linear regression smoothers and their minimax efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J., and I. Gijbels (1992): "Variable Bandwidth and local linear regression smoothers," *The Annals of Statistics*, 20, 2008–2036.

Gallant, A., and G. Tauchen (1996): "Which moments to match?," *Econometric Theory*, 12, 657–681.

Gasser, T., and H. Müller (1979): *Smoothing Techniques for Curve Estimation Kernel estimation of regression funtions*pp. 23–68.

Hall, P., S. Nath Lahiri, and J. Polzehl (1995): "On bandwidth choice in nonparametric regression with both short and long-range dependent errors," *The Annals of Statistics*, 23, 1921–1936.

Hansen, L., and J. A. Scheinkman (1995): "Back to the future: Generating moment implications for continuous–time Markox processes," *Econometrica*, 63, 767–804.

Litterman, R., and J. Scheinkman (1991): "Common Factors affecting Bond Returns," *Journal of Fixed Income*, 1, 54–61.

Lo, A. W. (1988): "Maximum likelihood estimation of Generalized Ito Processes with discretely-sampled data," *Econometric Theory*, 4, 231–247.

Longstaff, F., and E. Schwartz (1992): "Interest rate volatility and the term structure: A two-factor general equilibrium model," *Journal of Finance*, 47, 1259–1282.

Nadaraya, E. (1964): "On estimating regression," *Theory of Probability and its Applications*, 9, 141–142.

Oksendal, B. (1995): *Stochastic Differential Equations*. Springer, Berlin.

Pearson, N., and T. Sun (1994): "Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model," *Journal of Finance*, 49, 1279–1304.

Priestley, M., and M. Chao (1972): "Nonparametric function fitting," *Journal of Royal Statistical Society*, 34, 385–392.

Rudemo, M. (1982): "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, 9, 65–78.

Ruppert, D., and M. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22,3, 1346–1370.

Schaefer, S., and E. Schwartz (1984): "A Two-Factor Model of Term Structure: An Approximate Analytical Solution," *Journal of Financial and Quantitative Analysis*, 19, 413–424.

Scott, D. (1992): *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York.

Stanton, R. (1997): "A Non parametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, 52, 1973–2002.

Wand, M., and M. Jones (1995): *Kernel Smoothing*. Chapman and Hall.

Watson, G. (1964): "Smooth regression analysis," *Sankhya Series*, 26, 359–372.

Appendix A

This appendix is devoted to state some basic definition about the kernel density estimators and derive the Asymptotic Mean Integrated Squared Error (MISE) in order to compute the optimal bandwidth matrix $\mathbf{H}$.

$(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ denote a d-variate random sample having density $f(\cdot)$. I will use the notation $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})'$ to denote the components of $\mathbf{X}_i$ and a generic vector $\mathbf{x} \in \mathbb{R}^d$ will have the representation $\mathbf{x} = (x_1, \ldots, x_d)$. Also $\int$ will be the shorthand for $\int \ldots \int_{\mathbb{R}^d}$.

A d-dimensional kernel density estimator is

$$\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i) \tag{35}$$

where $\mathbf{H}$ is a symmetric positive semidefinite $d \times d$ matrix called the *bandwidth matrix* and

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

is a d-variate kernel function. Usually, it is a function satisfying:

1. $\int K(\mathbf{x}) d\mathbf{x} = \mathbf{1}$
2. $\int \mathbf{x}\mathbf{x}^{\mathbf{T}} K(\mathbf{x}) d\mathbf{x} = \mu_2(K)\mathbf{I}_d$
3. $\int \mathbf{x} K(\mathbf{x}) d\mathbf{x} = \mathbf{0}$

where $\mu_2(K) = \int x_i^2 K(\mathbf{x}) d\mathbf{x}$ is independent of i.

The kernel function is often taken to be a d-variate probability density function. A common technique for generating multivariate kernels from a symmetric univariate kernel k is the *product kernel*:

$$K^P(\mathbf{x}) = \prod_{i=1}^{d} k(x_i)$$

The analysis of the performance of the kernel density estimator requires the specification of appropriate errors criteria. In classical parametric statistics it is common to measure the closeness of an estimator using the Mean Squared Error (MSE). In a density estimation framework the MSE is rewrite in the following way:

$$MSE(\hat{f}(\mathbf{x}; \mathbf{H})) = \mathrm{E}[\hat{f}(\mathbf{x}; \mathbf{H}) - f(\mathbf{x})]^2 = \mathrm{Var}(\hat{f}(\mathbf{x}; \mathbf{H})) + \ \mathrm{Bias}^2(\hat{f}(\mathbf{x}; \mathbf{H}))$$

where Bias $= \left[\mathrm{E}(\hat{f}(\mathbf{x};\mathbf{H})) - f(\mathbf{x})\right]$. The disadvantage with MSE is that it evaluate the distance between $\hat{f}(\cdot)$ and $f(\cdot)$. at some point of the support. From a data analytic view point, to estimate f over the entire line is necessary an error criteria that *globally* measure the distances between the function $\hat{f}(\mathbf{x};\mathbf{H})$ and $f(\mathbf{x})$. One criterion is the Integrated Squared Error (ISE) (or the square of the L_2 distance):

$$ISE = \int (\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x}))^2 dx$$

A improvement of ISE is MISE (Mean Integrated Squared Error) that is the expected value of ISE. It has the advantage to consider not only the dataset at hand but take into account other possible datasets from the density f:

$$\begin{aligned} MISE &= \mathrm{E}\left(\int (\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x}))^2 dx\right) = \int MSE(\hat{f}(\mathbf{x};\mathbf{H})) = \\ &= \int \mathrm{Var}(\hat{f}(\mathbf{x};\mathbf{H})) + \int \mathrm{Bias}^2(\hat{f}(\mathbf{x};\mathbf{H})) \end{aligned}$$

If we substitute $\hat{f}(\mathbf{x};\mathbf{H})$ with eq. (35) the result is a function of the convolution of $K_{\mathbf{H}}$ with the unknown density function $f(\cdot)$ (see Wand and Jones (1995)). Using some "tractable" kernel density function, (i.e. the d-variate normal density) is possible to simplify the equation but it remain a function of bandwidth $\mathbf{H}$ in a complicate way.

The solution to this problem is to approximate asymptotically MISE (AMISE) , using multivariate version of the Taylor's formula.

Let $\mathcal{F}$ denote the class of symmetric positive definite $d \times d$ matrices. In general $\mathbf{H}$ has $\frac{1}{2}d(d+1)$ independent entries which, even for moderate d, can be a substantial number of smoothing parameters to have to choose.

A simplification can be obtained by imposing the restriction $\mathbf{H} \in \mathcal{D}$, where $\mathcal{D} \subset \mathcal{F}$ is the subclass of diagonal positive definite $d \times d$ matrices: $\mathbf{H} = \mathrm{diag}(h_1^2, \ldots, h_d^2)$.

For $\mathbf{H} \in \mathcal{D}$ the kernel estimator can then be written:

$$\hat{f}(\mathbf{x};h) = n^{-1}\left(\prod_{l=1}^{d} h_l\right)^{-1} \sum_{i=1}^{n} K(\frac{x_1 - X_{i1}}{h_1}, \cdots, \frac{x_d - X_{id}}{h_d})$$

A further simplification follows from the restriction $\mathbf{H} \in \mathcal{S}$ where $\mathcal{S} = \{h^2\mathbf{I} : h > 0\} \subset \mathcal{D}$ and leads to the single bandwidth kernel estimator:

$$\hat{f}(\mathbf{x};h) = n^{-1}h^{-d}\sum_{i=1}^{n} K\{(\mathbf{x} - \mathbf{X_i})/h\}$$

Following the same decomposition of MISE, the AMISE can be written as the sum of asymptotic integrated square bias and asymptotic integrated variance. To evaluate the asymptotic square bias is necessary to approximate via multivariate taylor's theorem the expected value of $\hat{f}(\mathbf{x};\mathbf{H})$:

$$\mathrm{E}(\hat{f}(\mathbf{x};\mathbf{H})) = \int K_{\mathbf{H}}(\mathbf{x}-\mathbf{u})f(\mathbf{u})d\mathbf{u} = \int K(\mathbf{z})f(\mathbf{x}-\mathbf{H}^{\mathbf{1/2}}\mathbf{z})d\mathbf{z}$$

Let introduce the Taylor series expansion of second order of $f(\mathbf{x}-\mathbf{H}^{\mathbf{1/2}}\mathbf{z})$ around $\mathbf{x}$ where ∇ is the gradient and $\mathcal{H}$ is the hessian:

$$\begin{aligned}
\mathrm{E}(\hat{f}(\mathbf{x};\mathbf{H})) &\approx \int K(\mathbf{z})\{f(\mathbf{x}) - (\mathbf{H}^{\mathbf{1/2}}\mathbf{z})'\nabla(\mathbf{x}) + 1/2((\mathbf{H}^{\mathbf{1/2}}\mathbf{z})'\mathcal{H}(\mathbf{H}^{\mathbf{1/2}}\mathbf{z}))\}d\mathbf{z} \\
&+ o(tr(\mathbf{H})) \\
&= f(\mathbf{x}) - \int \mathbf{z}'\mathbf{H}^{\mathbf{1/2}}\nabla(\mathbf{x})K(\mathbf{z})d\mathbf{z} + \frac{1}{2}\int \mathbf{z}'\mathbf{H}^{\mathbf{1/2}}\mathcal{H}\mathbf{H}^{\mathbf{1/2}}\mathbf{z}K(\mathbf{z})d\mathbf{z} \\
&+ o(tr(\mathbf{H}))
\end{aligned}$$

Using the property (1) of $K(\mathbf{x})$, and some manipulations of the *trace* (tr) operator, is possible to show that the leading bias term is:

$$\mathrm{E}(\hat{f}(\mathbf{x};\mathbf{H})) - f(\mathbf{x}) \approx \frac{1}{2}\mu_2(K)tr\{\mathbf{H}\mathcal{H}(\mathbf{x})\}$$

The variance of $\hat{f}(\mathbf{x};\mathbf{H})$is given by:

$$\mathrm{Var}(\hat{f}(\mathbf{x};\mathbf{H})) = -n^{-1}|\mathbf{H}|^{-1/2}R(K)f(\mathbf{x}) + o(n^{-1}|\mathbf{H}|^{-1/2})$$

where $R(K) = \int K(\mathbf{z})^2 d\mathbf{z}$.

Putting all together:

$$AMISE(\hat{f}(\mathbf{x};\mathbf{H})) = -n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}\mu_2(K)^2\int tr^2\{\mathbf{H}\mathcal{H}(\mathbf{x})\}d\mathbf{x}$$

It's possible to simplify the expression if $\mathbf{H} \in \mathcal{S}$:

$$AMISE(\hat{f}(\mathbf{x};\mathbf{H})) = -n^{-1}h^{-d}R(K) + \frac{1}{4}h^4\mu_2(K)^2\int \mathcal{I}^2 d\mathbf{x}$$

where

$$\mathcal{I}^2 = \sum_{i=1}^{d}(\partial^2/\partial x_i^2)f(\mathbf{x})$$

In order to obtain the optimal bandwidth h is necessary to minimize AMISE with respect to h. Equating the derivative of AMISE respect to h to zero and solving for h is possibile to obtain:

$$h_{AMISE} = \left[\frac{dR(K)}{\mu_2(K)^2 \int \mathcal{I}^2 d\mathbf{x} n}\right]^{1/(d+4)} = Cn^{-1/(d+4)} \tag{36}$$

where $C = \left[\frac{dR(K)}{\mu_2(K)^2 \int \mathcal{I}^2 d\mathbf{x}}\right]^{1/(d+4)}$ can be defined as a "scaling factor"[4]

One drawback in using the h_{AMISE} is that it depend on the unknown derivatives of the true density function $f(\cdot)$. In many case there are no possibilities to know what is the density that generate the data and so h_{AMISE} in useless. This is the reason for the data-driven bandwidth selectors (or "hi-tech" bandwidth selectors).

One of these is the *Least Square Cross Validation* (LSCV) (see Bowman (1984) and Rudemo (1982)) that is derived from MISE:

$$\hat{\mathbf{H}}_{LSCV} = \arg\min_{\mathbf{H}\in\mathcal{F}} LSCV(\mathbf{H})$$

$$LSCV(\mathbf{H}) = \int \hat{f}(\mathbf{x};\mathbf{H})^2 - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(\mathbf{X}_i;\mathbf{H})$$

where $\hat{f}_{-i}(\mathbf{X}_i;\mathbf{H})$ is the kernel estimator based on the sample with $\mathbf{X}_i$ deleted. Of course, one can also minimise LSCV over $\mathbf{H} \in \mathcal{D}$ or $\mathbf{H} \in \mathcal{S}$ to obtain bandwidth selectors belonging to a smaller class. In this paper the LSCV is applied with $\mathbf{H} \in \mathcal{D}$.

[4]Even if $\mathbf{H} \in \mathcal{D}$ is possible to show that $\mathbf{H}_{AMISE}$ is a diagonal matrix and each diagonal element is the bandwidth of i-th ($1 \leq i \leq d$) random variable. Each of them has a representation like eq. (36)

Appendix B

Regression model is a natural setup to estimate conditional moments. There are several approaches to the nonparametric regression problem: kernel functions, spline functions and wavelets. In the context of kernel regression traditional approaches have involved the Nadaraya-Watson estimator (Nadaraya (1964), Watson (1964)) and some other alternative kernel estimators (Priestley and Chao (1972), Gasser and Müller (1979)). For example Stanton (1997) use the Nadaraya-Watson kernel regression in order to estimate conditional moments. In this paper a different class of kernel-type regression estimators is used: the local polynomial kernel estimators. These estimate the regression function at a particular point by "locally fitting" a pth degree polynomial to the data via weighted least squares, where the weights are chosen according to the height of a kernel function centered about that point. This class includes, as a special case, the Nadaraya-Watson estimator since it can be shown to correspond to fitting degree zero polynomials.

Of particular importance is the local linear kernel estimator, corresponding to $p = 1$. The local linear kernel estimator shares some similarities with the above mentioned traditional kernel regression estimators, although it has good asymptotic properties and most important, favourable boundary behaviour compared with those. The last is the main reason for its use in this paper.

A regression model start with the usual equation:

$$\mathbf{Y} = m(\mathbf{X}) + \epsilon$$

where $\mathbf{Y}$ is the response variable, $\mathbf{X}$ are the regressors, $m(\cdot)$ is the regression function and ϵ is the error variable.

In a nonparametric setting, the functional form of $m(\cdot)$ is not specified and only mild conditions are required (i.e. smoothness condition and existence, at least, of first and second derivatives in order to derive asymptotic MISE). The goal is to estimate $m(\mathbf{X})$ since that:

$$m(\mathbf{x}) = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$$

is the conditional moment used to approximate drift, diffusion and correlation function of diffusion process (i.e. in the paper, $\mathbf{Y}$ is substituted with the changes of the daily rates, and the regressors $\mathbf{X}$ are the conditioning variables: the short and spread rate). A local polynomial kernel estimator of $m(\cdot)$,$\hat{m}(\cdot)$, is a function of the point $\mathbf{x}$ at which we evaluate the regression function, the degree of the polynomial p, and the bandwidth matrix $\mathbf{H}$ used to evaluate the optimal weights to assign at the sample points close the support point $\mathbf{x}$.

Using a notation as much as possible similar to appendix A, the estimator $\hat{m}(\mathbf{x}, p, \mathbf{H})$ is obtained by fitting the polynomial:

$$\beta_0 + \beta_1(\mathbf{x} - \cdot) + \ldots + \beta_p(\mathbf{x} - \cdot)^p$$

to the $(\mathbf{X}_i, \mathbf{Y}_i)$ using weighted least squares with kernel weights $K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$. The value of $\hat{m}(\mathbf{x}, p, \mathbf{H})$ is the height of the fit $\hat{\beta}_0$ where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ minimizes:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i}^{n} \{Y_i - \beta_0 - \cdots - \beta_p(\mathbf{x} - \mathbf{X}_i)^p\}^2 K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

Assuming the invertibility of $\mathbf{X}_{\mathbf{x}}'\mathbf{W}_{\mathbf{x}}\mathbf{X}_{\mathbf{x}}$, standard weighted least squares theory leads to the solution

$$\hat{\beta} = (\mathbf{X}_{\mathbf{x}}'\mathbf{W}_{\mathbf{x}}'\mathbf{X}_{\mathbf{x}})^{-1}\mathbf{X}_{\mathbf{x}}'\mathbf{W}_{\mathbf{x}}\mathbf{Y}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of responses,

$$\mathbf{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{x} - \mathbf{X}_1) & \cdots & (\mathbf{x} - \mathbf{X}_1)^p \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ 1 & (\mathbf{x} - \mathbf{X}_n) & \cdots & (\mathbf{x} - \mathbf{X}_n)^p \end{pmatrix}$$

is an $n \times (p+1)$ matrix and

$$\mathbf{W}_{\mathbf{x}} = \text{diag}\ (K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_1), \ldots, K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_n))$$

is an $n \times n$ diagonal matrix of weights.

Since the estimator of $m(\mathbf{X})$ is the intercept coefficient we obtain:

$$\hat{m}(\mathbf{x}, p, \mathbf{H}) = \mathbf{e}_1'(\mathbf{X}_{\mathbf{x}}'\mathbf{W}_{\mathbf{x}}'\mathbf{X}_{\mathbf{x}})^{-1}\mathbf{X}_{\mathbf{x}}'\mathbf{W}_{\mathbf{x}}\mathbf{Y}$$

where $\mathbf{e}_1$ is a $(p+1) \times 1$ vector having 1 in the first entry and zero elsewhere.

Fan (1993) and Fan and Gijbels (1992), in a univariate setting, show that the local linear kernel regression estimator has asymptotic properties making it superior to the Nadaraya-Watson and Gasser-Müller kernel estimator.

Ruppert and Wand (1994) have derived asymptotic bias and variance to the case of multivariate predictor variables. The authors show that, assuming some conditions, the asymptotic bias in an interior point of the support of the density function $f(\mathbf{x})$ is:

$$\mathrm{E}\{\hat{m}(\mathbf{x}; \mathbf{H}) - m(\mathbf{x}) | \mathbf{X}_1, \ldots, \mathbf{X}_n\} = \frac{1}{2}\mu_2(K) tr\{\mathbf{H}\mathcal{H}_m(\mathbf{x})\} + o_p\{tr(\mathbf{H})\}$$

with $R(K) = \int K(\mathbf{u})^2 d\mathbf{u}$, $\mathcal{H}_m$ is the $d \times d$ Hessian matrix of a sufficiently smooth d-variate function m at $\mathbf{x}$.

The variance of $\hat{m}(\mathbf{x}; \mathbf{H})$ is given by:

$$\mathrm{Var}\{\hat{m}(\mathbf{x}; \mathbf{H})|\mathbf{X}_1, \ldots, \mathbf{X}_n\} = [n^{-1}|\mathbf{H}|^{-1/2} R(K)/f(\mathbf{x})] v(\mathbf{x})[1 + o_p(1)]$$

where $v(\mathbf{x}) = \mathrm{Var}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$

Putting all together to determine AMISE:

$$\begin{aligned} AMISE(\hat{m}(\mathbf{x}; \mathbf{H})|\mathbf{X}_1, \ldots, \mathbf{X}_n) &= \int [n^{-1}|\mathbf{H}|^{-1/2} R(K)/f(\mathbf{x})] v(\mathbf{x})[1 + o_p(1)] d\mathbf{x} + \\ &+ \int \left(\frac{1}{2} \mu_2(K) tr[\mathbf{H} \mathcal{H}_m(\mathbf{x})] + o_p[tr(\mathbf{H})] \right)^2 d\mathbf{x} \end{aligned}$$

It's possible to simplify the expression if $\mathbf{H} \in \mathcal{S}$:

$$\begin{aligned} AMISE(\hat{m}(\mathbf{x}; \mathbf{H})|\mathbf{X}_1, \ldots, \mathbf{X}_n) &= n^{-1} h^{-d} R(K) \int \frac{v(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x} + \\ &+ \frac{1}{4} h^4 \mu_2(K)^2 \int \mathcal{I}^2 d\mathbf{x} \end{aligned}$$

where

$$\mathcal{I}^2 = \sum_{i=1}^{d} (\partial^2 / \partial x_i^2) f(\mathbf{x})$$

In order to obtain the optimal bandwidth h i follow the same procedure of appendix A about kernel density estimation and the solution is:

$$h_{AMISE} = \left[\frac{dR(K) \int \frac{v(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x}}{\mu_2(K)^2 \int \mathcal{I}^2 d\mathbf{x} n} \right]^{1/(d+4)} = C n^{-1/(d+4)} \tag{37}$$

where $C = \left[\frac{dR(K)}{\mu_2(K)^2 \int \mathcal{I}^2 d\mathbf{x}} \right]^{1/(d+4)}$ can be defined as a "scaling factor" in an analogue way of kernel density estimation.

The problem using this formula is the evaluation of the scaling factor that depend on unknown functions (i.e. density function and the variance). One way could be the data-driven bandwidth selectors and more specific the cross validation bandwidth selector. The goal is to minimize some "estimated"

mean square error criteria that are function of the bandwidth: the Average Squared Error (ASE) or its expectation (MASE), the MISE or AMISE.

Frequently cross validation procedure is performed minimizing, with respect to $\mathbf{H}$, the Estimated Integrated Squared Error (EISE):

$$\arg\min_{\mathbf{H}} CV(\mathbf{H}) = \arg\min_{\mathbf{H}} n^{-1} \sum (\mathbf{Y} - m(\mathbf{x}; \mathbf{H}))^2$$

Considering eq. (37) the optimization has been conducted with a search over a grid of bandwidth values in order to arrive to an optimal bandwidth. In this paper $\mathbf{H}$ is a 2×2 diagonal matrix ($\mathbf{H} = \text{diag}(h_1^2, h_2^2)$ with $d = 2$) and h_1, h_2 are respectively the bandwidth for the short and spread rate. The idea is to fix a matrix of possible "scaling factors", one for the short and spread rate (the only unknown value of eq. (37)). The choice has been a square matrix $\mathbf{C}$ of scaling factors of dimension 12×12 with $c_{1,\cdot}, c_{\cdot,2} = \{1, 2, \ldots, 12\}$. For each c_{ij} has been computed h_1, h_2 (according to eq. (37)), the corresponding bandwidth matrix $\mathbf{H}$ and $CV(\cdot)$ function. At the end we have a 12×12 matrix of $CV(\cdot)$ values. The minimum value has been chosen and so the optimal $\hat{\mathbf{H}}$ has been found. The procedure is computational slow and an alternative optimization strategy has been adopted subject to the hypothesis that the $CV_{\mathbf{H}}(\cdot)$ function is sufficiently smooth.

Instead of calculate the $CV_{\mathbf{H}}(\cdot)$ function for each entries of the $\mathbf{C}$ matrix, the algorithm fix the value of c along one dimension (i.e. the first value of $c_{\cdot,2}$ corresponding to h_2) and calculate the bandwidth matrix for different value of c along the second dimension (i.e. different values of h_1). For each value of $\mathbf{H}$, has been computed the $CV_{\mathbf{H}}(\cdot)$ function. The first step of the strategy is completed when $CV_{\mathbf{H}}(\cdot)$ function reach the minimum value or the difference between to consecutive values are less than a tolerance threshold defined by the researcher.

The second step start fixing the value of h_1 at the level reached in the previous step and calculating $CV_{\mathbf{H}}(\cdot)$ function for each $\mathbf{H}$ bandwidth matrix obtained changing the value of h_2 (due to a change in $c_{\cdot,2}$) that in the first step was fixed. This step terminate when the same conditions of the previous one are reached.

The third and last step start with the value of $\mathbf{H}$ determined before (that is the result of the levels reached by h_1 and h_2) and creating a sub-square grid of 5×5 around the value of $\mathbf{H}$ (or better the value c_{ij}) that represent its neighborhood in two dimension. For each entries of this sub-square grid has been calculated the corresponding bandwidth matrix $\mathbf{H}$ and $CV_{\mathbf{H}}(\cdot)$ function. The step terminate if the minimum value of $CV_{\mathbf{H}}(\cdot)$ correspond to the center of the sub-square grid. If not, another sub-square grid is determined around the new minimum and the procedure continue until minimum is the center

of the sub-square grid. For the dataset at hand this strategy reach the minimum of $CV_{\mathbf{H}}$ avoiding to compute each entry of $\mathbf{C}$ matrix and it's three times faster.

The cross validation function has been calculated using a Jackknife-based procedure or what is called in nonparametric theory the "leave-k-out" cross validation where k is equal to 100. Chu and Marron (1991) consider this modified form of cross-validation to take into account short-range dependence in the data. k has been fixed to 100 after some trials with different value of k (see Hall, Nath Lahiri, and Polzehl (1995) for a discussion) .

Appendix C

In this appendix are derived the first, second and third order approximations of μ_R, μ_S, σ_R, σ_S, σ_{RS}.

For $N = 1$ the approximation is already derived in eq. (30). Substituting $f(\mathbf{X}_t)$ with the functions in eq.(32) and using the calculation of eq. (33) is possible to write the following approximations:

$$\begin{aligned}
\mu_R(R_t, S_t) &= \frac{1}{\Delta}\mathrm{E}_t(R_{t+\Delta} - R_t) + O(\Delta) \\
\mu_S(R_t, S_t) &= \frac{1}{\Delta}\mathrm{E}_t(S_{t+\Delta} - S_t) + O(\Delta) \\
\sigma_R^2(R_t, S_t) &= \frac{1}{\Delta}\mathrm{E}_t[(R_{t+\Delta} - R_t)^2] + O(\Delta) \\
\sigma_S^2(R_t, S_t) &= \frac{1}{\Delta}\mathrm{E}_t[(S_{t+\Delta} - S_t)^2] + O(\Delta) \\
\sigma_{RS}(R_t, S_t) &= \frac{1}{\Delta}\mathrm{E}_t[(R_{t+\Delta} - R_t)(S_{t+\Delta} - S_t)] + O(\Delta)]
\end{aligned}$$

For $N = 2$, the linear combination of $\hat{\mathrm{E}}^i(\cdot)$ becomes:

$$\alpha_1\hat{E}(\mathbf{X}_t) + \alpha_2\hat{E}(\mathbf{X}_t) = [\alpha_1 + \alpha_2]\mathcal{L}f(\mathbf{X}_t) + \frac{1}{2}[\alpha_1 + \alpha_2 2]\mathcal{L}^2 f(\mathbf{X}_t)\Delta$$

$$[\alpha_1 + \alpha_2]\mathcal{L}f(\mathbf{X}_t) = \alpha_1\hat{E}(\mathbf{X}_t) + \alpha_2\hat{E}(\mathbf{X}_t) - \frac{1}{2}[\alpha_1 + \alpha_2 2]\mathcal{L}^2 f(\mathbf{X}_t)\Delta$$

The value of α_1, α_2 will be chosen in order to sum to 1 and to eliminate the term of order $O(\Delta)$. This restriction are summarized in the following expression:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\alpha \equiv V\alpha = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \; ; \alpha = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

Doing the several mentioned substitutions the approximations of second order are:

$$\begin{aligned}
\mu_R(R_t, S_t) &= \frac{1}{2\Delta}[4\mathrm{E}_t(R_{t+\Delta} - R_t) - \mathrm{E}_t(R_{t+2\Delta} - R_t)] + O(\Delta^2) \\
\mu_S(R_t, S_t) &= \frac{1}{2\Delta}[4\mathrm{E}_t(S_{t+\Delta} - S_t) - \mathrm{E}_t(S_{t+2\Delta} - S_t)] + O(\Delta^2)
\end{aligned}$$

$$
\begin{aligned}
\sigma_R^2(R_t, S_t) &= \frac{1}{2\Delta}\left[4\mathrm{E}_t[(R_{t+\Delta} - R_t)^2] - \mathrm{E}_t[(R_{t+2\Delta} - R_t)^2]\right] + O(\Delta^2) \\
\sigma_S^2(R_t, S_t) &= \frac{1}{2\Delta}\left[4\mathrm{E}_t[(S_{t+\Delta} - S_t)^2] - \mathrm{E}_t[(S_{t+2\Delta} - S_t)^2]\right] + O(\Delta^2) \\
\sigma_{RS}(R_t, S_t) &= \frac{1}{2\Delta}[4\mathrm{E}_t[(R_{t+\Delta} - R_t)(S_{t+\Delta} - S_t)] - \\
&- \mathrm{E}_t[(R_{t+2\Delta} - R_t)(S_{t+2\Delta} - S_t)]] + O(\Delta^2)
\end{aligned}
$$

For $N = 3$ the Vandermond and vector α becomes:

$$
\alpha = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ 1 \end{pmatrix}
$$

and the approximations are:

$$
\begin{aligned}
\mu_R(R_t, S_t) &= \frac{1}{6\Delta}[18\mathrm{E}_t(R_{t+\Delta} - R_t) - 9\mathrm{E}_t(R_{t+2\Delta} - R_t) + 2\mathrm{E}_t(R_{t+3\Delta} - R_t)] + O(\Delta^3) \\
\mu_S(R_t, S_t) &= \frac{1}{6\Delta}[18\mathrm{E}_t(S_{t+\Delta} - S_t) - 9\mathrm{E}_t(S_{t+2\Delta} - S_t) + 2\mathrm{E}_t(S_{t+3\Delta} - S_t)] + O(\Delta^3) \\
\sigma_R^2(R_t, S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(R_{t+\Delta} - R_t)^2] - 9\mathrm{E}_t[(R_{t+2\Delta} - R_t)^2] + 2\mathrm{E}_t[(R_{t+3\Delta} - R_t)^2]) + O(\Delta^3) \\
\sigma_S^2(R_t, S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(S_{t+\Delta} - S_t)^2] - 9\mathrm{E}_t[(S_{t+2\Delta} - S_t)^2] + 2\mathrm{E}_t[(S_{t+3\Delta} - S_t)^2]) + O(\Delta^3) \\
\sigma_{RS}(R_t, S_t) &= \frac{1}{6\Delta}(18\mathrm{E}_t[(R_{t+\Delta} - R_t)(S_{t+\Delta} - S_t)] - 9\mathrm{E}_t[(R_{t+2\Delta} - R_t)(S_{t+2\Delta} - S_t)]) + \\
&+ 2\mathrm{E}_t[(R_{t+3\Delta} - R_t)(S_{t+3\Delta} - S_t)] + O(\Delta^3)
\end{aligned}
$$

Figure 1— Plots of Short Rate and Spread

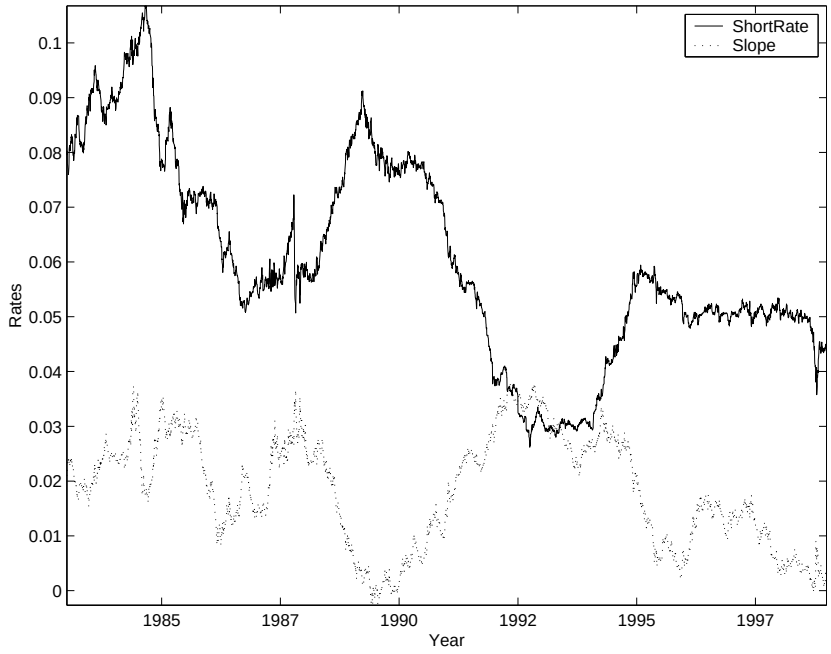

Figure 2— Scatter plot of Short Rate vs Spread

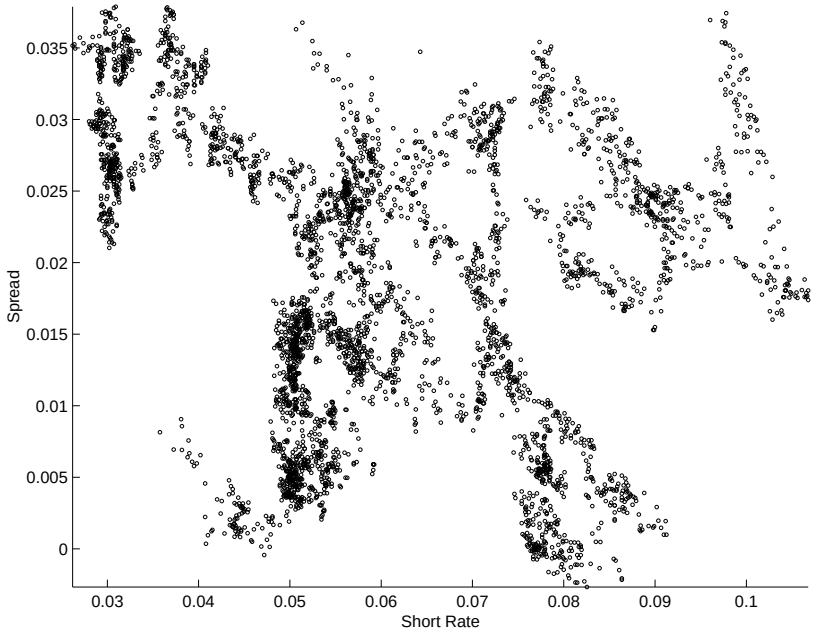

Figure 3— Histograms Short and Spread rate with superimposed fitted normal density

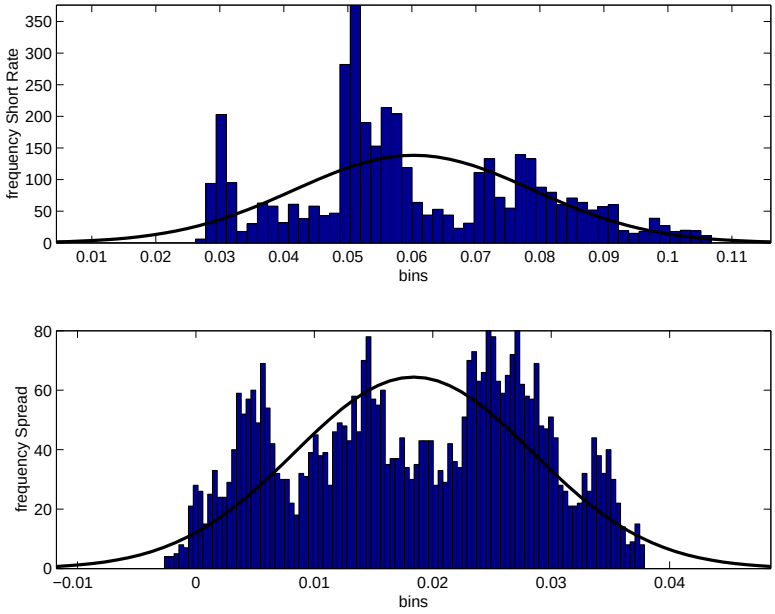

Table 1— Descriptive statistics Short rate (level rate: r_t)

N	Mean	Median	Max	Min
4004	6.027	5.640	10.677	2.619
Stand. Dev.	Skewness	Kurtosis	Jarque Bera	p value JB
1.860	0.283	2.366	120.506	0.000

Table 2— Descriptive statistics Spread rate (level rate: r_t)

N	Mean	Median	Max	Min
4004	1.839	1.907	3.788	-0.266
Stand. Dev.	Skewness	Kurtosis	Jarque Bera	p value JB
1.005	-0.126	1.916	206.544	0.000

Table 3— Descriptive statistics Short rate (first difference: $r_t - r_{t-1}$)

N	Mean	Median	Max	Min
4003	-0.000889	0.0000	0.452000	-0.541000
Stand. Dev.	Skewness	Kurtosis	Jarque Bera	p value JB
0.061766	-0.306298	11.81619	13026.50	0.000

Table 4— Descriptive statistics Spread rate (first difference: $r_t - r_{t-1}$)

N	Mean	Median	Max	Min
4003	-0.000477	0.000000	0.468000	-0.326000
Stand. Dev.	Skewness	Kurtosis	Jarque Bera	p value JB
0.066592	0.324651	7.262630	3100.926	0.000

Table 5— Autocorrelation Short rate
(level r_t and first difference: $r_t - r_{t-1}$)

	ρ_1	ρ_3	ρ_5	ρ_7
r_t	0.999	0.997	0.996	0.994
$r_t - r_{t-1}$	0.086	-0.038	0.019	0.013

Table 6— Autocorrelation Spread rate
(level r_t and first difference: $r_t - r_{t-1}$)

	ρ_1	ρ_3	ρ_5	ρ_7
r_t	0.997	0.992	0.988	0.983
$r_t - r_{t-1}$	0.060	-0.063	0.047	-0.005

Table 7— Augmented Dickey-Fuller stationary test
for first difference of short rate

H_0	Test statistic	Critical Value (10%)
Nonstationary	-27.95896	-2.5675

Table 8— Augmented Dickey-Fuller stationary test for
first difference of spread rate

H_0	Test statistic	Critical Value (10%)
Nonstationary	-28.98246	-2.5675

Figure 4— Nonparametric density estimation Short rate

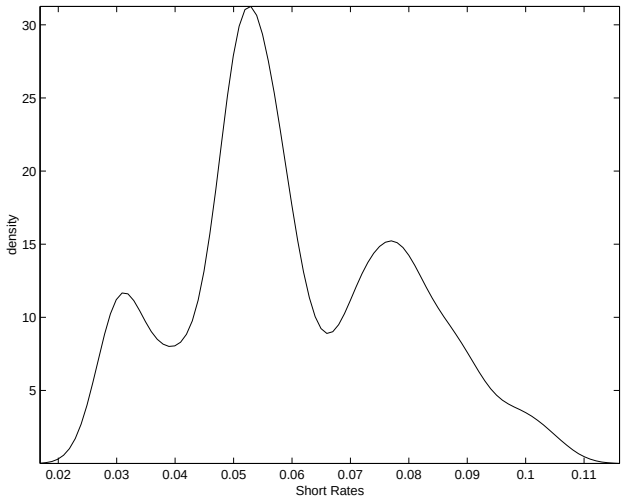

Figure 5— Nonparametric density estimation Spread rate

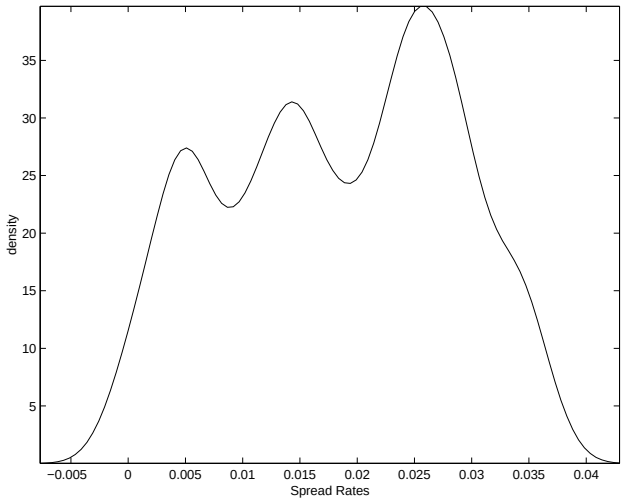

Figure 6— Nonparametric bivariate density estimation Short and Spread rates

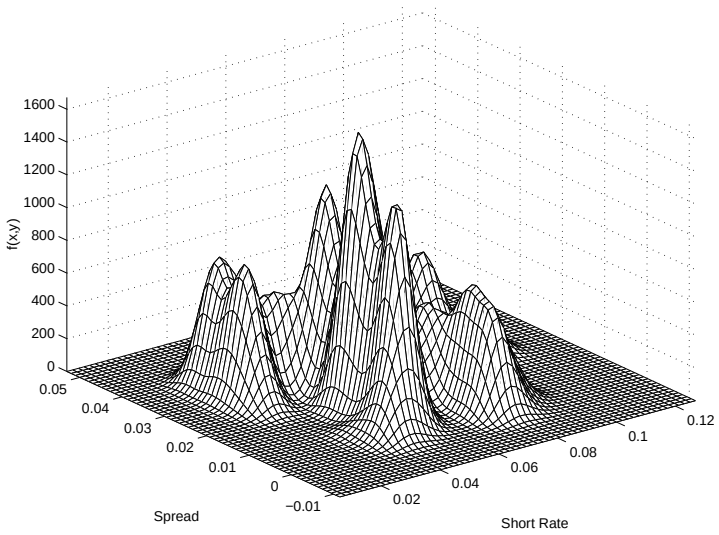

Figure 7— Nonparametric bivariate density Short and Spread rates (other perspective)

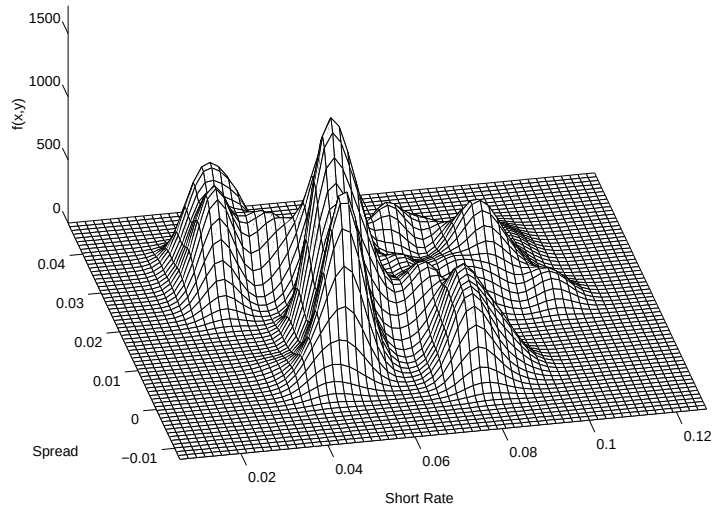

Figure 8— Third order approximation to the drift of the Short rate, conditional on Short and Spread rate

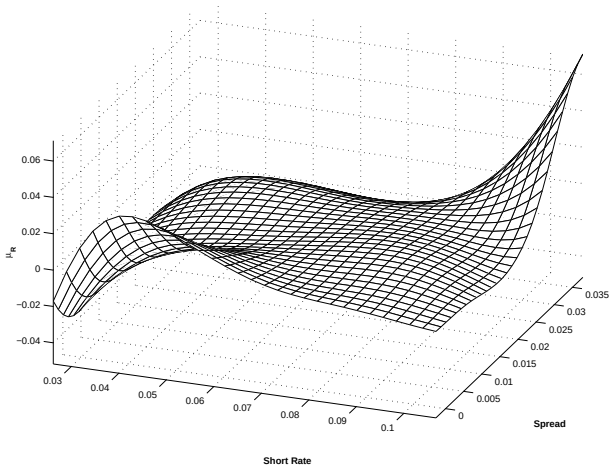

Figure 9— Third order approximation to the diffusion of the Short rate, conditional on Short and Spread rate

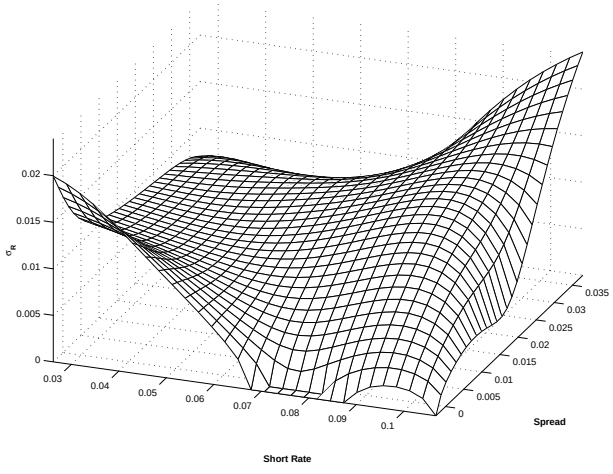

Figure 10— Third order approximation to the drift of the Spread rate, conditional on Short and Spread rate

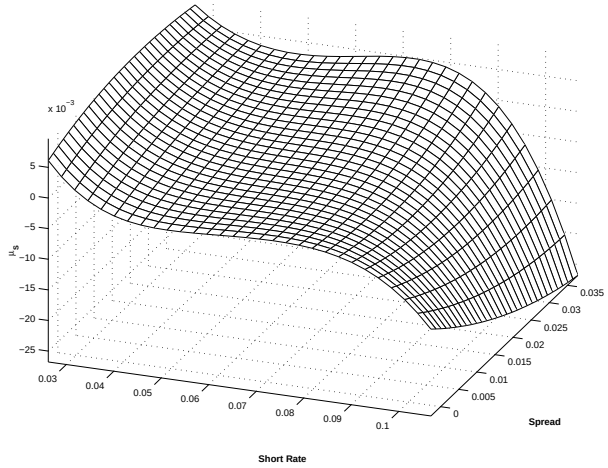

Figure 11— Third order approximation to the diffusion of the Spread rate, conditional on Short and Spread rate

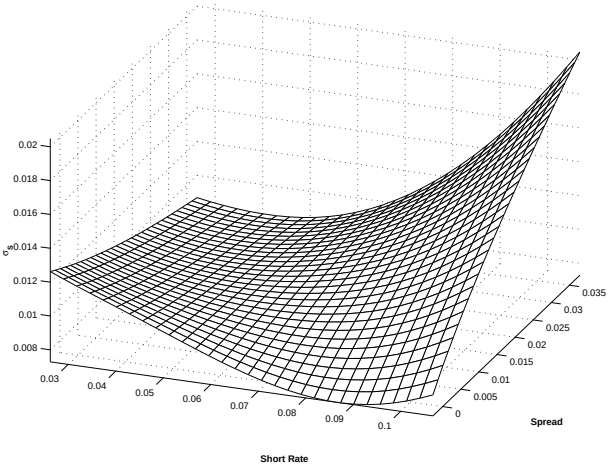

Figure 12— Third order approximation to the correlation between Short and Spread rate, conditional on Short and Spread rate